U.S. DEPARTMENT OF COMMERCE
Patent and Trademark Office

DOCUMENT RETRIEVAL REQUEST FORM

Requester's Name: Uyen Le	Case Serial Number: 9/947221		Art Unit/Org.: 2100
Phone: 2-4021	RightFax:	Building:	Room Number: 3D21
Date of Request: 11/30/04		Date Needed By: RUSH	

Paste or add text of citation or bibliography: **Paste Citation** Only one request per form. Original copy only. ☐

Author/Editor:	See Attached
Journal / Book Title	
Article Title	
Volume Number:	Report Number: Pages:
Issue Number:	Series Number: Year of Publication:
Publisher:	
Remarks: (178)	520102

Library Action	PTO		LC		NAL		NIH		NLM		NIST		Other	
	1st	2nd	1st	2nd	1st	2nd	1st	2nd	1st	2nd	1st	2nd	1st	2nd
Local Attempts	X													
Date	11/30/04													
Initials	SG													
Results	NA													
Examiner Called														
Page Count														
Money Spent														

		Source	Date
Remarks/Comments: 1st and 2nd denotes time taken to a library O/N – Under NLM means Overnight	Ordered From:	CAST ordered & Complete Faxed	
	Comments:		

USCOM-DC

PTO-XXXX (2-98)

mount of stored data. This paper introduces a **data mining** system is introduced. The main feature of this system is specially design fuzzy rule induction algorithm which extracts useful pattern in databases. Since fuzzy logic has the **affinity** with the human knowledge representation, it considered as an essential component of **data mining** systems.
Descriptors: *Fuzzy sets; Data processing; Database systems; Data storage equipment; Statistics; Data reduction; Knowledge based systems; Neural networks; Relational database systems; Backpropagation
Identifiers: **Data mining** ; Fuzzy rule induction; Database interface; Verification module
Classification Codes:
723.4.1 (Expert Systems)
921.4 (Combinatorial Mathematics, Includes Graph Theory, Set Theory); 723.2 (Data Processing); 722.1 (Data Storage, Equipment & Techniques); 922.2 (Mathematical Statistics); 723.4 (Artificial Intelligence)
921 (Applied Mathematics); 723 (Computer Software); 722 (Computer Hardware); 922 (Statistical Methods)
92 (ENGINEERING MATHEMATICS); 72 (COMPUTERS & DATA PROCESSING)

for 08/947,221
U. Le
24021

46/5/3 (Item 3 from file: 8)
DIALOG(R)File 8:Ei Compendex(R)
(c) 2003 Elsevier Eng. Info. Inc. All rts. reserv.

178

04002780 E.I. No: EIP94121505160
Title: Discovering interesting statements from a database
Author: Gebhardt, F.
Corporate Source: Gesellschaft fur Mathematik und Datenverarbeitung mbH (GMD), Sankt Augustin, Ger
Source: Applied Stochastic Models and Data Analysis v 10 n 1 Mar 1994. p 1-14
Publication Year: 1994
CODEN: ASMAEM ISSN: ~~0875-0024~~ 8755-0024
Language: English
Document Type: JA; (Journal Article) Treatment: T; (Theoretical)
Journal Announcement: 9502W1
Abstract: **Knowledge discovery** aims at extracting new knowledge from potentially large databases; this may be in the form of interesting statements about the data. Two interrelated classes of problem arise that are treated here: to put the subjective notion of 'interesting' into concrete terms and to deal with large numbers of statements that are related to one another (one rendering the other redundant or at least less interesting). Four increasingly subjective facets of 'interestingness' are identified: the subject field under consideretion, the conspicuousness of a finding, its novelty, and its deviation from prior knowledge. A procedure is proposed, and tried out on two quite different data sets, that allows for specifying interestingness by various means and that ranks the results in a way that takes interestingness (relevance, evidence) as well as mutual relatedness (similarity, **affinity**) into account - manifestations of the second and third facets of interestingness in the given data environment. (Author abstract) 14.
Descriptors: *Database systems; Knowledge based systems; Data reduction; Data acquisition; Data structures; User interfaces; Algorithms; Statistical methods
Identifiers: **Knowledge discovery** ; Exploratory data analysis; Interestingness; Statements; Data sets; Facets
Classification Codes:
723.4.1 (Expert Systems)
723.3 (Database Systems); 723.4 (Artificial Intelligence); 723.2 (Data Processing); 922.2 (Mathematical Statistics)
723 (Computer Software); 922 (Statistical Methods)
72 (COMPUTERS & DATA PROCESSING); 92 (ENGINEERING MATHEMATICS)

46/5/4 (Item 1 from file: 2)
DIALOG(R)File 2:INSPEC
(c) 2003 Institution of Electrical Engineers. All rts. reserv.

APPLIED STOCHASTIC MODELS AND DATA ANALYSIS, VOL. 10, 1–14 (1994)

DISCOVERING INTERESTING STATEMENTS FROM A DATABASE

F. GEBHARDT

Gesellschaft für Mathematik und Datenverarbeitung mbH (GMD), Schloß Birlinghoven, Postfach 13 16, D-53731 Sankt Augustin, Germany

SUMMARY

Knowledge discovery aims at extracting new knowledge from potentially large databases; this may be in the form of interesting statements about the data. Two interrelated classes of problem arise that are treated here: to put the subjective notion of 'interesting' into concrete terms and to deal with large numbers of statements that are related to one another (one rendering the other redundant or at least less interesting). Four increasingly subjective facets of 'interestingness' are identified: the subject field under consideration, the conspicuousness of a finding, its novelty, and its deviation from prior knowledge. A procedure is proposed, and tried out on two quite different data sets, that allows for specifying interestingness by various means and that ranks the results in a way that takes interestingness (relevance, evidence) as well as mutual relatedness (similarity, affinity) into account—manifestations of the second and third facets of interestingness in the given data environment.

KEY WORDS Knowledge discovery in databases Exploratory data analysis Interestingness Project EXPLORA

1. DISCOVERING STATEMENTS

During the past years, knowledge discovery in databases has attracted growing attention. The aim is to extract new knowledge from data sets in the form of hidden dependencies that may hold either in all cases or in a statistical sense; prior knowledge guides or supports the discovery process in various ways.

An example is project EXPLORA.[1,2] Statistical dependencies are found and presented to the user in the form of textual statements indicating, for example, subpopulations (groups defined by explanatory variables) that differ significantly from the overall average; on request, the text is augmented by the relevant figures. Background knowledge consists in this case mainly in the ordinal or hierarchical structure of variables and in the distinction between explanatory variables and variables to be explained. This results in presenting only the most comprehensive statements, suppressing subsets as redundant or uninteresting. EXPLORA is able to handle statements composed of one, two or more variables; thus the system has to construct vast search graphs and, once a significant statement has been found, to cut off its subgraphs. This is not a trivial task since sub-subgraphs may have many other ancestors.

While finding remarkable results quite rapidly, this concept suffers from some drawbacks. The subpopulation (group of objects) found by the search algorithm is formally significant by construction, but need not really be interesting; the real cause for the apparent significance

CCC 8755–0024/94/010001–14

Received 30 August 1992
Revised 6 August 1993

may be a subgroup which is highly significant while the rest of the subpopulation is more or less inconspicuous.

There may be a number of similar statements where the supporting groups are nevertheless not subsets of one another, such as 'income above 3000, age above 30', 'income above 3500, age above 25', 'income above 4000, age between 20 and 60'. All of these will be presented, and occasionally this amounted to about a dozen similar statements tending to confuse the user.

There may be several statements that are formally completely distinct but nevertheless express more or less the same constellation, the cause being strong relations between certain values of different variables like 'retired persons' and 'persons above 60'.

A solution to the first problem could be always to check the next lower level and to present the higher level only when the lower one is rather homogeneous; in cases of doubt, both levels could be shown. This would further increase the number of results.

Procedures of this kind reduce the mass of original data to a set of selected statements, possibly a large set so that the user wishes to reduce it further according to his (or her) interests. This leaves us with two sorts of problem.

The search algorithm may find several statements that in fact express the same constellation; selecting just the strongest result is inadequate since the strength is subject to statistical variation; there is a fluid transition from 'same constellation' to 'unrelated'. How should one select the most important results?

What does 'interesting' mean; which tools could enable the user to express personal, subjective directions of interest?

We shall identify some increasingly subjective facets of interestingness and then present an algorithm handling a set of related statements with the aim of emphasizing those that are presently the most interesting ones. Roughly speaking, a statement that appears to be less important is devalued and pushed down in the ordered list of noticeable statements. It does not get lost, but others are given priority.

2. INTERESTINGNESS

2.1 Broad and narrow meanings

The word 'interesting' is used in many papers, but the authors rarely say what they mean by it. If an explanation is given, it tends to signify the degree by which a result deviates from average or normal. Thus Piatetski-Shapiro[3] measures the interestingness of rule $A \rightarrow B$ discovered in a database essentially by the function

$$|A \cap B| - |A|\,|B|/|M|$$

after some normalization where A and B are two subpopulations of population M and $|\cdot|$ denotes as usual the size of a set.

EXPLORA in its present implementation utilizes a broader concept. Interestingness is based on the degree of deviation from average or normal, but redundant statements, i.e. statements following from another interesting statement, are considered uninteresting. The same holds true for statements that are to be expected once an interesting statement is presented; this usually means that statements concerning a subgroup are considered uninteresting if the corresponding statement on the group is already exhibited.

A short paper by Lebowitz[4] delves into the question of when a short newspaper notice is interesting. The goal is not to store uninteresting parts from the very beginning. He finds three

grounds for non-interest: (1) concepts not causally connected to the main point of the story, such as accidental details; (2) concepts that can be reconstructed, such as details usually connected to the main fact (their absence could be interesting); and (3) concepts that are overshadowed, such as the death of the driver if an important person has been shot.

Later on, Lebowitz[5] applies these ideas in the machine learning system UNIMEM combining explanation-based and similarity-based learning: the interest in certain features of a story partially guides the search process for generalizations. He stresses the heuristic dimension of what is interesting. The paper cites other sporadic attempts to put the notion of interest in concrete form.

But interestingness can be still more diffuse and rather intangible, even idiosyncratic. A story becomes interesting if one knows one of the acting persons. A telephone number is interesting to a mathematician if it starts with the same digits as the number π. Sometimes, the interesting point of a message (an election result, a hit list, a company telephone directory) is that an expected result or entry is missing.

Certainly, an automated knowledge discovery system cannot anticipate such individual or even queer connections that arouse interest; but one should stay aware that they exist. Any system supporting the user in finding interesting facts should therefore offer sufficient means for browsing and for directing the search.

We concentrate now on aspects of 'interesting' that are related to the goal of knowledge extraction from databases and are susceptible to some kind of formalization.

2.2. Aspects of interestingness

Given a big and unfamiliar database, what features or aspects or facets of interestingness can be utilized to guide or delimit the search for interesting pieces of knowledge?

Firstly, the *subject field under consideration* determines a broad boundary of what is interesting. The user specifies this aspect by choosing the proper database and probably a subset of the data such as certain groups of variables or objects. We include here also the (temporary) selection of a type of dependencies. In EXPLORA, this is defined by an uninitialized validation function and the choice of variables to be inserted into it. Connected to this function is the linguistic template for stating the results found by the search algorithm.

Secondly, the *conspicuousness or evidence of a finding* delimits the degree of interestingness. Thy result must be unusual, unexpected, or important according to some pre-defined criterion. We assume that the criterion not only determines whether a finding (a statement) is (true and) worth recording, but also yields a numerical value for the degree of conspicuousness. We shall see later that this first starting point may be modified in order to take other facets of interestingness into account.

The conspicuousness may, but need not, rest on a statistical significance measure; therefore we avoid the term 'significance'. In our examples, we try to characterize a given subset of a set of objects by the values of selected descriptive variables; the conspicuousness will be based on the sizes of the involved sets. There may be dozens of such characterizations of the given subset yielding similar values of conspicuousness.

Thirdly, the *novelty* or *dissimilarity* of a statement with respect to other results found at the same time influences the interestingness. If a statement on a group of objects is automatically or usually also valid for a subgroup, the latter one is redundant and therefore uninteresting. If two conspicuous groups overlap to a large extent, one of them could suffice for describing the situation. However, if the goal is to find an explanation, both statements should probably be considered.

We propose an algorithm ranking the statements in such a way that the conspicuousness as well as the dissimilarity to all previously exhibited statements is taken into account.

Fourthly, the *deviation from prior knowledge* is an important ingredient of interestingness. However, this aspect is hardly practicable. The computer cannot assess the user's prior knowledge. Even if it could, the latter might have changed: the user has learned some new facts and forgotten others. Moreover, suppressing an outstanding, but known, result could make the user believe it does not hold. We do not, therefore, consider this aspect further.

These aspects are not independent of each other. For instance, prior knowledge could be incorporated into the evaluation function. In that case the user knows what assumptions are already taken into consideration and wrong conclusions should be avoided.

It is clear that interestingness is a highly subjective matter. Different persons will almost certainly disagree on the relative interestingness of a bunch of statements. We nevertheless propose a procedure for extracting interesting results, but there are two safeguards: the algorithm provides parameters for individual tuning, and no conspicuous statements become lost in an absolute sense, they are merely rearranged in a rank order originating from that procedure.

2.3. Test data

The ideas presented in this paper have been extensively tried out on two test data sets: election data and financial services data.

For the Federal elections in 1987, Germany was subdivided into 248 constituencies. According to law, these have to have approximately the same number of persons entitled to vote; the existing differences in size have therefore been neglected.

There were four major parties, CDU (in Bavaria CSU), SPD, FDP and 'Die Grünen'. Their election results as well as the differences from the previous election (in 1983) serve as the variables in which the user is interested, for which he wants to find conspicuous results. More specifically, one strives to charactelize the 50 (or 30 or 62, for example) best or worst election districts by one or two of the ten demographic variables.

Most variables including the election outcomes are originally given as percentages; examples are portions of unemployed, of persons employed in agriculture, of persons above 65 years of age; the exceptions are Bundesland (State) and population density.

For each demographic variable (except Bundesland) some class boundaries have been set in advance such as 11·5%, 12%, 12·5% and 13% for young persons (18 to 25 years). Only the sets of constituencies falling below or above such a boundary have been used in describing the best or worst districts of a party. Thus a sample statement for the best 62 districts of SPD (election result) is 'Agriculture $\leqslant 5\%$, unemployment $> 12\%$ (18 districts, 16 of them belonging to the 62 best ones)'.

The second data set is a survey on 20000 people regarding their affiliation with bank services. The demographic variables include hierarchic ones such as geographical region and occupation, numerical ones such as income and age, other ordinal variables such as school education and nominal ones such as sex and marital status.

2.4. Examples for the evidence

Two quite different evidence measures have been used based on the following contingency

table:

	Goal objects	Non-goal objects
Group	n_{11}	n_{12}
Complement	n_{21}	n_{22}

When one is looking for groups with a high proportion of goal objects, the basis for the evidence is $n_{11}/n_{11}+n_{12}$). The maximal value is 1; it is also attained for a group consisting of a single goal object. To give small groups a disadvantage, our first evidence function is

$$V_1 = n_{11}/(2+n_{11}+n_{12})$$

In addition, we restricted the search to groups with at least 8 or 10 objects (constituencies). This evidence favours groups with a high portion of goal objects; these groups are usually small, the best ones containing mostly ten to twenty constituencies.

The second evidence function is the usual χ^2. Let e_{ij} be the expected value of cell (i,j) given the row and column sums; then

$$V_2 = \chi^2 \approx \Sigma\ (n_{ij} - e_{ij})^2/e_{ij}$$

This function favours large groups since they are statistically more significant. A good typical statement (50 goal objects) refers to a group with 40 to 50 objects, about three quarters of them being goal objects.

The second data set (financial services) has been explored using V_2. Since the data are confidential, no concrete examples can be communicated, but in general the behaviour resembled that of the election data (processed with V_2). Owing to the large sample size, the values of χ^2 were larger; but otherwise the differences between the dependent variables (between parties or between financial services) were more pronounced than those between the two data sets, e.g. regarding the distance between the first evidences or the number of statements with an evidence at least a certain percentage of that of the best one.

3. RANK ORDER

3.1. Sorting according to the evidence

From information retrieval, it is well known that sorting the hits—if there are more than a dozen or so—according to a suitable rank order has a big advantage to the user. Nothing gets lost, but there is a chance that the first few documents already answer the posed question. The ranking algorithms are based on the frequencies of the occurrences of the search terms in the individual documents. Surprisingly, few commercial systems offer this feature.

In the case of acquiring statistically confirmed knowledge from databases it should also be helpful to sort the results according to their strengths. Sometimes the task leads naturally to a grouping; in this case, the results should be sorted according to strength within each group. If one is looking for conspicuous election results, a grouping according to parties suggests itself.

There is hardly a natural boundary between relevant statements and irrelevant ones. Conventional significance boundaries are of little value since one performs conceptionally perhaps tens of thousands of tests that in addition, are not independent of each other; therefore, a formal significance boundary of 0·1 % or less is at best a very rough guide.

Throughout the paper we assume that data exploration yields so many results that it is a

problem to extract the really interesting ones. The first thing to do is then to sort the search results according to the evidence expressed by the evaluation function, possibly within groups provided by the nature of the problem. In what follows we investigate refinements that better reflect the interest of the user.

3.2. Actual variations to the evidence

The evaluation function—perhaps several functions to be used at the discretion of the user—is installed in the EXPLORA knowledge discovery system either from the very beginning or when a particular database is initiated; it reflects the anticipated general interests of all users.

However, for any session the user may have varying interests. As a partial remedy we now propose that the system provides means for selectively modifying the initial value of the evidence. This is a means for treating the second aspect of interestingness listed in Section 2.2, conspicuousness or evidence of a finding.

The user should certainly have a tool for deleting individual resulting statements that at present are of no interest. But this is not the point here. What is needed are means for increasing or decreasing the initial evidence for specified types of results corresponding to facets of interest. This can be done by adding or subtracting a fixed constant or by multiplying with a constant near 1, this choice depending on the nature of the evidence.

In order to have a reasonable effect, the constant must be big enough (or differ sufficiently from 1) to change the order of statements considerably, but small enough not to dominate the new order; the selected group of statements should in general neither displace all others nor be pushed completely out of sight.

No search result gets lost, only the order is altered. This lessens the danger of perhaps unwillingly creating the outcome one is eager to verify, i.e. to lie with statistics.

We now give examples of groups of statements for which the discovery system could provide tools to revalue or devalue the evidence. We assume an EXPLORA-like setting where statements describe conspicuous groups (subpopulations) of the population under examination.

The user modifies all statements that contain a particular variable. The reason for devaluation could be that this variable is known to be unreliable or expensive to obtain, or, more likely, the result rather than the cause of what is to be explained; similarly for revaluation.

Statements containing just one variable are considered more valuable; therefore more complex statements are degraded according to the number of variables.

Higher levels of a hierarchic variable are preferred; therefore the evidence of statements containing lower levels is decreased.

For a metric or ordinal variable, interior intervals (intervals having a lower and an upper boundary) are more complicated and usually more difficult to interpret than open intervals; therefore they are degraded.

Sometimes there exists a temporal dimension in the data: some statements pertain to more recent facts than others or some statements in a time-varying data set have gained importance since a specified date, perhaps the day when this user last attempted to discover novelties. In such a context newer results can be upgraded.

A database allows diverse types of evaluations, such as finding groups with positive and negative deviations from the average or, alternatively, considering groups deviating at a certain

time or relative to a previous date. These can of course be considered separately, but, if the statements are mixed, it may be desirable to favour one type over the other.

Interest being as variable as it is, there is probably no formalized way to test these ideas, but they have mostly been tried out on the two databases described in Section 2.3. The author admits being biased, but it seems that these modifications of the original evidence work out quite well and show the expected behaviour.

The exact value of the changes to the evidence are not critical; according to our experiences it is advisable to use the same constant with all facets of interest and to provide for means to upgrade or degrade a statement or a type of statements more than once.

For clarity, we repeat that these re- and devaluations reflect the momentary interests of the user; thus they cannot be incorporated into the original evidence.

4. DEVALUATION DUE TO AFFINITY

4.1 Affinity of statements

We have already mentioned that a statement may be uninteresting because another one which is quite similar is already known or accepted; the interest focuses on novel and dissimilar results as introduced in the third aspect of interestingness in Section 2.2.

The word 'similarity' is often used in connection with a distance obeying the triangular inequality, which is not assumed here to hold; we therefore prefer 'affinity'.

We assume now that for any pair of statements A_i and A_j an affinity $S(A_i, A_j)$ or $S(i,j)$ for short is given. This affinity is to be normalized to $0 \leqslant S \leqslant 1$. $S(A_i, A_j) = 0$ means there is no relationship or affect; knowing A_j does not influence the interestingness of A_i. $S(A_i, A_j) = 1$ signifies the strongest influence of A_j on A_i possible. If in addition the evidence for A_i is considerably lower than that for A_j, A_i becomes uninteresting.

This formulation is by design asymmetric, for we do not even assume $S(A_i, A_j) = S(A_j, A_i)$. We now give two examples.

4.2. Examples for affinity

The first evidence function, V_1, aims at finding groups of objects (election districts in our case) with a high percentage of goal objects. Clearly, subgroups of such a group will in general also comprise a high portion of goal objects; thus, they are of little interest. On the other hand, a larger group can be interesting even if the portion of goal objects is smaller. This situation demands an asymmetric affinity. We chose

$$S_1(A_i, A_j) = |M_i \cap M_j| / |M_i|$$

where $|M|$ is a measure for the size of group M. If the user is looking for a description of the election results, $|M|$ could be the number of goal objects of M. Two groups containing the same goal objects have affinity one; it does not matter which one we take as far as the affinity is concerned: the decision rests on the evidence only. If, however, the user is searching for an explanation, both groups are of interest in the case where their non-goal objects are different even if one of them has a higher evidence.

In the case of the second evidence, V_2, it seems appropriate to base the affinity also on χ^2. The rows and columns of the contingency table now correspond to the two groups and their complements. Since S has to be normalized, we divide by the highest possible value, which is

N, (the total number of objects):

$$S_2 = \chi^2/N$$

If both groups coincide, then $S_2 = 1$.

If S is an affinity function, then so is S^x for $x > 0$. Small values of x emphasize any similarities that are present, while large values restrict the influence of the affinity to groups that are equal or nearly equal.

In a similar context (Section 5 of Reference 14) it has been shown that one should use S_2 with $x < 0{\cdot}5$. This recommendation is based on an analysis of the case when one group is a subgroup of the other; then the larger one should be uninteresting if its deviation from the overall mean results purely from the deviation of the subgroup.

4.3. Devaluation algorithm

We shall return now to the task of presenting to the user those statements which are supposedly most interesting.

In principle, all conspicuous statements are shown, but the most interesting ones should be shown first. However, the user is free to stop the presentation when a chosen limit is reached.

The primary order is given by the evidence. But this order has to be altered: if statement A_j depends highly on statement A_i but has a markedly lower evidence, then it is much less interesting than other statements with evidences similar to $V(A_j)$ but smaller relationships (affinities) to A_i. If however $V(A_i)$ and $V(A_j)$ are (nearly) equal, the two statements should not affect one another.

Roughly speaking, statement A_i reduces the effective evidence of statement A_j by an amount that depends on the affinity and on the distance between the evidences. One has to make sure, however, that a statement A_i that has thus been heavily devalued does not influence another whose evidence is well below $V(A_i)$ but above the devalued evidence for A_i. This suggests the following rules, which may at first seem somewhat complicated but are not, as the subsequent algorithm shows.

We introduce restricted evidences $R_i(A_j)$ and $R(A_j)$ by means of

$$R_i(A_j) = V(A_j) \cdot [V(A_j)/R(A_i)]^{\delta S(j,i)}$$
$$R(A_j) = \min_i R_i(A_j)$$

with a free parameter δ to be chosen by the user. R_i is the result of A_i devaluing $V(A_j)$; the final restricted evidence is the minimum with respect to i. If for some i the square brackets have value greater than 1, this i cannot yield the minimum and can be ignored from the beginning.

The algorithm for computing the restricted evidences proceeds in four steps.

1. For all statements A_j, let $R(A_j) = V(A_j)$.
2. Among all statements not yet presented to the user, let A_i be that one with the highest reduced evidence $R(A_i)$. It is now presented to the user.
3. For the remaining statements, the reduced evidence is updated: $R(A_j) \leftarrow \min(R(A_j), R_i(A_j))$.
4. Return to step 2 unless all statements have been presented or a user defined stop criterion has been reached.

Iteration.

In this algorithm, the user's interests can be quantified in two ways.

Choosing an exponent x with the affinity determines what constitutes a high affinity. Assume

a total of 248 objects, as in the election data, and two groups of 20 objects each with 15 objects in common. This yields $S_2 = 0{\cdot}53$. If one considers these groups to be rather similar, one could use $x = 0{\cdot}5$, leading to the effective affinity $S_2^{\frac{1}{2}} = 0{\cdot}73$. If one wants by and large to disregard such situations, $x = 2$ results in $S_2^2 = 0{\cdot}28$. This cuts the influence of one group on the other in half; but both operations have little effect on affinities close to 1.

The second choice for the user is δ; δ determines how far one statement devalues another that is highly related (affinity close to 1). δ should be large enough to produce a striking effect except perhaps in the first half dozen statements (these very often have quite similar evidences such that the devaluation remains small). In our test data, the range from 1 to 4 turned out to be a good choice.

Two things should be stressed again. The exact values of x and δ are not important; changing them slightly will switch here and there the order of two statements, but nothing becomes lost. The whole procedure might look rather like 'lying with statistics'; but we are here in the context of exploratory statistics and knowledge discovery and the aim is not to prove something statistically but to find relationships that are interesting enough for further study and confirmation.

4.4. Limiting behaviour

In order to get an impression of the results that can be expected, we shall discuss the behaviour of the devaluation algorithm for extreme values of δ and x. We assume without loss of generality that the statements are indexed according to their evidences, i.e. $V(A_i) \geq V(A_{i+1})$ for all i. To simplify the discussion, we assume furthermore that all evidences are different, i.e. $V(A_i) > V(A_{i+1})$, unless indicated otherwise.

Let δ be fixed and assume $0 < S(S_i, A_j) < 1$ for all i and j $(i \neq j)$. If $x \to 0$, $S(A_i, A_j)^x \to 1$. Then the devaluation of A_j is determined by the largest quotient $V(A_j)/R(A_i)$; this is clearly $V(A_j)/V(A_1)$ since A_1 is not devalued at all and all other original evidences—and therefore all reduced evidences—are smaller than, or at most equal to, $V(A_1)$. Thus all statements are devalued according to their quotients $V(A_j)/V(A_1)$, and that leaves their order unchanged.

If for some j, $S(A_1, A_j) = 0$, then A_j is not devalued by A_1, but it may of course be devalued by some other A_i. However, we then have

$$V(A_j)/R(A_i) < V(A_j)/V(A_1)$$

so that A_j is devalued less than statements with similar evidence but positive affinity to A_1. Therefore A_j can advance to an earlier position; of course it depends on the other evidences and affinities as to which place it will occupy.

Now let $x \to \infty$, implying $S(A_i, A_j)^x \to 0$. Then no devaluation takes place; the order is again unchanged. If, however, for some i and j, $V(A_i) > V(A_j)$ and $S(A_i, A_j) = 1$, then statement A_j is devalued. Thus statements that have affinity 1 with another statement with higher evidence can be pushed down in the list in this limiting case.

In summary, only certain statements are affected with regard to their order in the limiting cases $x \to 0$ and $x \to \infty$, but in between the order may change severely. There is however no way to predict which value of x has the largest effect; in fact, as the test data have shown, this can be different with similar data such as different dependent variables within the same data set using the same type of evidence and affinity and the same independent variables.

Next let x be fixed and δ vary. For $\delta = 0$, the evidence is unchanged: $R(A_i) = V(A_j)$. To show the limiting behaviour for $\delta \to \infty$, let us take logarithms of the definition of $R_j(A_j)$:

$$\log R_i(A_j) = \log V(A_j) + \delta S(A_j, A_i) \log [V(A_j)/R(A_i)]$$

If A_j is devalued by A_i at all (i.e. $S(A_j, A_i) \neq 0$ and $V(A_j)/R(A_i) < 1$), then the second term (which is negative at least for $i = 1$) essentially determines R_j. Thus the final order is that given by

$$S(A_j, A_1) \log [V(A_j)/R(A_1)]$$

with some obvious modifications if there exist more statements with the same evidence as A_1.

If j is the smallest index such that $S(A_j, A_1) = 0$, then A_j is not devalued by A_1 and for δ large enough, all other statements will be devalued below $R(A_j)$. Thus the final order is given by A_1, then all statements with affinity 0 with A_1 and between one another and finally all other statements ordered according to

$$S(A_j, A_1) \log [V(A_j)/R(A_1)]$$

In our test data, one needs unreasonably large values for δ to approach this limiting order for the first few dozen statements, often $\delta > 100$ or even $\delta > 1000$ for useful values of x. But this analysis demonstrates the tendency of the effect of large δ to favour statements that are (nearly) unrelated to the first statement and to each other.

4.5. Sample results

As a demonstration of the procedure, we look at the 30 best constituencies of the party 'Die Grünen' according to the percentage of votes in 1987 (this means at least 11·2%). In this case we used the evidence function V_1 emphasizing groups with a high percentage of goal objects; we restricted the search to groups containing at least 8 out of the 30 districts. The affinity function for this example was S_1 based on all objects.

Among the groups that turned out to be of interest are those shown in Table I.

The meaning of the descriptive variables is as follow:

Cath	proportion of the population who are Roman Catholics
Dens	population density (inhabitants per square kilometre)
Emp	proportion of the population who are employed
Old	proportion of the population who are above 65 years of age
Prod	proportion of the working population who are in productive trades (industry)
Serv	proportion of the working population who are in service trades (%); highly correlated with Prod (Prod, Serv and agriculture add up to 100%)
Unempl	proportion of persons willing to work who are unemployed (%)
Young	proportion of the population who are of age 18 to 25 years (%).

Some descriptions cover identical districts: groups 6 and 7, 9 and 10, 13 and 14, 17 and 18. This is due to the high correlation between Prod and Serv; Serv > 60 (46 objects, 24 goal objects) is a subset of Prod ≤ 40 (49 objects, 24 goal objects).

Table II shows the first ten groups that are offered by the devaluation algorithm with various values of the free parameters x and δ. It is not recommended to use x as low as 0·1 or as high as 20, or $\delta = 8$; these lines are added to demonstrate what happens in the limit.

Groups 56, 63, 92, 101 and 126 have no elements in common with group 1 (nor with groups 2 to 5). In addition, groups 56 and 63 are fairly different (only 11 districts in common), while 92, 101 and 126 are similar to 56. The dissimilarity to groups 1 and 2 is of course the reason they show up at low levels of x. The surprising property of groups 63 and 101 is that they have a *high* unemployment rate. They are, however, almost subsets of group 6.

$\delta = 1$ shows little effect (as was to be expected). With $\delta = 2$ and $\delta = 4$, group 2 maintains its

place due to the small difference in evidence to group 1. Groups 3 to 5 are also quite similar to group 1, but their lower evidences result in stronger devalutions. Groups 17 and 18 are much larger than group 1 and less similar to group 1 than, for example, group 15; therefore they are only slightly devalued and show up at early places at several parameter combinations.

Next let us look at the same data with a different affinity function, S_1 based on goal elements only, i.e. $|M|$ is the number of goal elements in group M. Table III shows the first ten statements offered by the devaluation algorithm for $\varkappa = 0{\cdot}5$, 1, 2 and 4 and $\delta = 1$, 2 and 4.

At first glance, the differences to Table II are only minor. Some statements appear at somewhat earlier places, in particular groups 56 and 63. Actually, their affinity to group 1 is 0 both times, so they have not changed. But some other groups have been devalued more heavily.

A good example is group 9. Five of its objects also belong to group 1; all five are goal objects. Thus $S_1(A_9, A_1) = 1/3$ when based on all objects and, 1/2 when based on the goal objects.

Similarly, $S_1(A_6, A_1)$ rises from 0·3 to 0·42. When A_6 is only slightly devalued, it devalues in turn group 56 (affinity 0·87 and 1·0, respectively); but as soon as it falls below $V(A_{56})$, group 56 remains unaffected, which is more often the case in Table III than in Table II. Remember that A_{56} has no elements in common with A_1 to A_5.

The procedure has been tried out with various combinations of the 30 or 50 or 62 best or worst constituencies based either on the percentages of votes or the gains and losses of the four major parties. Rarely does a statement with a number above 100 appear on the first ten places

Table I. Interesting groups of constituencies. Goal objects are the best 30 constituencies of 'Die Grünen' in the Federal elections of 1987

No.	Description	Objects in group	Goal objects in group	Evidence
1	Prod ⩽ 45, Unempl ⩽ 6	12	10	0·714
2	Serv > 55, Unempl ⩽ 6	11	9	0·692
3	Old > 23·5, Unempl ⩽ 6	12	9	0·643
4	Empl > 45, Unempl ⩽ 6	14	10	0·625
5	Dens > 1000, Unempl ⩽ 6	13	9	0·600
6	Empl > 45, Prod ⩽ 40	30	19	0·594
7	Empl > 45, Serv > 60	30	19	0·594
8	Dens > 2000, Empl > 45	25	16	0·593
9	Cath > 60, Prod ⩽ 40	15	10	0·588
10	Cath > 60, Serv ⩽ 60	15	10	0·588
12	Serv > 50, Unempl ⩽ 6	15	10	0·588
13	Old > 23·5, Prod < 40	27	17	0·586
14	Old > 23·5, Serv > 60	27	17	0·586
15	Empl > 45, Serv > 55	34	21	0·583
16	Young ⩽ 12, Serv > 60	29	18	0·581
17	Dens > 400, Prod ⩽ 40	36	22	0·579
18	Dens > 400, Serv > 60	36	22	0·579
29	Dens > 20, Serv > 60	41	24	0·558
56	Cath ⩽ 20, Empl > 45	15	9	0·529
63	Empl > 45, Unempl > 10	17	10	0·526
92	Dens > 1000, Cath ⩽ 20	16	9	0·500
101	Empl > 45, Unempl > 8	20	11	0·500
126	Empl > 40, Cath ⩽ 20	17	9	0·474

Table II. Identification numbers of groups in the order produced by the devaluation algorithm for various values of x and δ. The affinity S_1 is based on all objects

x	δ	First 10 groups									
0·1	1	1	2	3	4	56	63	5	6	7	8
0·1	2	1	2	56	3	63	4	101	92	5	6
0·1	4	1	2	56	63	3	101	92	4	126	6
0·1	8	1	2	56	63	101	92	3	4	126	6
0·5	1	1	2	3	4	56	6	7	63	8	9
0·5	2	1	2	56	63	3	4	6	7	8	101
0·5	4	1	2	56	63	3	101	92	6	7	4
0·5	8	1	2	56	63	101	92	3	6	7	4
1	1	1	2	3	6	7	4	8	17	18	16
1	2	1	2	56	63	3	6	7	17	18	16
1	4	1	2	56	63	6	7	17	18	3	16
1	8	1	2	56	63	101	6	7	92	17	18
2	1	1	2	3	6	7	8	17	18	16	15
2	2	1	2	6	7	17	18	16	15	8	13
2	4	1	2	56	6	7	63	17	18	16	15
2	8	1	2	56	63	17	18	6	7	16	29
4	1	1	2	6	7	8	9	10	13	14	3
4	2	1	2	6	7	8	15	13	14	9	10
4	4	1	2	6	7	17	18	16	15	8	13
4	8	1	6	7	17	18	16	15	13	14	2
20	1	1	2	6	7	8	9	10	15	3	13
20	2	1	2	6	7	8	9	10	15	17	18
20	4	1	2	6	7	8	9	10	15	17	18
20	8	1	6	7	8	9	10	17	18	19	15

Table II. Identification numbers of groups in the order produced by the devaluation algorithm for various values of x and δ. The affinity S_1 is based on the goal objects

x	δ	First 10 groups									
0·5	1	1	2	3	4	56	6	7	63	8	9
0·5	2	1	2	56	63	3	4	6	7	8	101
0·5	4	1	2	56	63	3	101	92	6	7	4
1	1	1	2	3	6	7	4	8	17	18	16
1	2	1	2	56	63	3	6	7	17	18	16
1	4	1	2	56	63	6	7	17	18	3	16
2	1	1	2	3	6	7	8	17	18	16	15
2	2	1	2	6	7	17	18	16	15	8	13
2	4	1	2	56	6	7	63	17	18	16	15
4	1	1	2	6	7	8	9	10	13	14	3
4	2	1	2	6	7	8	15	13	14	9	10
4	4	1	2	6	7	17	18	16	15	8	13

as is the case above. Usually small changes occur in the first four to eight places, but quite often one or two statements with numbers between 20 and 50 show up among the first ten. Similar experiences have been gained with the election data using V_2 and S_2 as well as with the financial data.

4.6. Actual variations to the affinity

Just as can the evidence, the affinity can be modified to reflect the actual interests of the user. It seems, however, that alterations are less beneficial here.

With our test data, only one situation arose where alterations to the affinity helped clarify the analysis. The financial data included an income variable with quite a number of partitions; all income intervals were candidates for statements or constituents thereof. Thus, some dependent variables produced many results containing income and another variable, e.g. education.

This annoyance can be reduced if all affinities between statements using the same variables are increased, e.g. replacing $S(A_i, A_j)$ by $S(A_i, A_j)^{1/2}$ or even by $S(A_i, A_j)^{1/4}$.

Another potential application is the mixture of two or more types of statement, in the simplest case describing groups where the dependent variable is above or below average. Here the affinities within a type could be increased (or those between types decreased, which in conjunction with a suitable new x amounts to the same thing).

5. CONCLUSIONS

Exploratory data analysis tries to boil huge data sets down to a number of statements concerning data of potential interest to the user. There already exists, of course, a diversity of methods for detecting interesting features in large data sets; these extend from elementary tools in exploratory data analysis[6] like stem-and-leaf displays and boxplots over sophisticated procedures for uncovering low-dimensional nonlinearities in high-dimensional data clouds (projection pursuits)[7,8] to specialized algorithms for detecting numerical laws of quite complex strucure[9,10] as well as other machine learning methods.[11,12]

Here, we have been concerned with the situation in which a search algorithm like that of EXPLORA has extracted a number of potentially noteworthy statements but this set is still too large to detect easily the really interesting results. In further reducing this set—or rather in ranking it—we had to reconcile the two conflicting aspects of interestingness, viz. conspicuousness and novelty.

The solution presented above draws on two ideas.

Firstly, statements that are quite similar to a better one are devalued, i.e. pushed further down on the list of results. In order to do so, one needs a measure for the importance or worth of a statement, its evidence, and a measure for 'similar', the affinity. The devaluation algorithm then solves the task.

Secondly, the user should be enabled to cope with varying needs and individualistic, subjective facets of interestingness. This is accomplished by the choice of free parameters (x and δ) and by temporarily altering, if necessary, individual evidences, groups of evidences or groups of affinities.

The obvious objection to such a procedure is that it involves the danger of manipulating the results. Therefore we stress again that we are here in the context of exploration. The tests are in no way statistically valid: the test procedure will usually be quite rough and the immense

number of conceptually performed tests renders all nominal significance levels obsolete anyway. As a result, a meaningful procedure for stressing certain statements and lowering the weight of others should be beneficial rather than damaging.

Extensive trials with two quite different data sets have in the author's opinion verified the viability and usefulness of the proposed procedure. Unfortunately, resources did not permit the validation of the procedures with external users.

The selection procedure has been adapted to the problem of selecting interesting regression models from the set of all submodels of the full regression model.[13] In this case, the conspicuousness rests on the opposing criteria of high multiple correlation and low number of independent variables (sparcity of the model), the affinity is determined by the variables that two models have in common.

REFERENCES

1. P. Hoschka, and W. Klösgen, 'A support system for interpreting statistical data', in G. Piatetsky-Shapiro and W. Frawley (eds), *Knowledge Discovery in Databases*, AAAI, Menlo Park, California, 1991, 325–345.
2. W. Klösgen, 'Problems for knowledge discovery in databases and their treatment in the statistics interpreter Explora', *International Journal for Intelligent Systems*, 7, 649–673 (1992).
3. G. Piatetsky-Shapiro, 'Discovery, analysis, and presentation of strong rules', in G. Piatetsky-Shapiro and W. Frawley (eds), *Knowledge Discovery in Databases*, AAAI, Menlo Park, California, 1991, pp. 229–248.
4. M. Lebowitz, 'Cancelled due to lack of interest', in Ann Drinan (ed.), *Proceedings of the 7th International Joint Conference on Artificial Intelligence (IJCAI)*, Vancouver, August 1981, IJCAI, Menlo Park, California, pp. 13–15.
5. M. Lebowitz, 'Integrated learning: controlling explanation', *Cognitive Science*, 10, 219–240 (1986).
6. J. W. Tukey, *Exploratory Data Analysis*, Addison-Wesley, Reading, Massachusetts, 1977.
7. J. H. Friedman, 'Exploratory projection pursuit', *Journal of the American Statistical Association*, 82, 249–266 (1987).
8. P. J. Huber, 'Projection pursuit', *Annals of Statistics* 13, 435–475 (1985).
9. S. Kocabas, 'Computational models of scientific discovery', *Knowledge Engineering Review*, 6, 259–305 (1991).
10. P. Langley, H. A. Simon, G. L. Bradshaw, and J. M. Zytkow, *Scientific Discovery*, MIT Press, Cambridge, Massachusetts, 1987, 357pp.
11. I. Kodratoff and R. S. Michalski (eds), *Machine Learning: an Artificial Intelligence Approach*, Vol III, Morgan Kaufman, San Mateo, 1990, 825pp.
12. G. Piatetsky-Shapiro and W. Frawley (eds), *Knowledge Discovery in Databases*, AAAI, Menlo Park, California, 1991, 525pp.
13. F. Gebhardt, 'Selecting regression models in exploratory data analysis', in R. Gutiérrez (ed.), M. J. Valderrama (ed.), *Applied Stochastic Models and Data Analysis: Proceedings of the 5th International Symposium on ASMDA*, Granada, April 1991, World Scientific, Singapore, 1991, pp. 262–273.
14. F. Gebhardt, 'Choosing among competing generalizations', *Knowledge Acquisition*, 3, 361–380 (1991).